[0002]

[Description of the Related Art:]

There is a method in which the results of the total number of times words appear is referenced, and words whose frequency of appearance exceed a set value are extracted from among the cut out words, then the correlation between the extracted words is evaluated and a cluster of co-occurring words, whose correlation is greater than a set value, is created. When doing this, by creating a category dictionary that corresponds to the text that is the object of analysis, it is possible to display the analysis results of that text.   (For example, refer to patent document 1.)

[0003]

[Patent Document 1]

Japanese patent Publication No. 2001-101194 (Fig. 1)

[0004]

[Problem to be Solved By the Invention]

However, there was a problem in that it was not possible to detect the characteristics of the text data based on the correlation between keywords extracted from the text data.

[0005]

The present invention takes into consideration the condition described above and makes it possible to detect the characteristics of the text data based on the correlation between keywords extracted from the text data.

[0006]

[Means for Solving the Problem]

The information-processing apparatus of claim 1 comprises: an input unit that inputs text data; a text-data-memory unit that stores text data; a word-cutting unit that executes a word-cutting process on text data; a syntax-analysis unit that performs a syntax-analysis process on the text data on which the word-cutting process was performed; a thesaurus-creation

unit that creates thesauruses from the text data on which the syntax-analysis process was performed; a thesaurus-memory unit that stores the thesauruses created by the thesaurus-creation unit; a thesaurus-sorting unit that performs a sorting process on the text data on which word-cutting and syntax-analysis were performed; a sorting-results-memory unit that stores the sorting results from the thesaurus-sorting unit; a frequency-of-appearance-calculation unit that calculates the frequency of appearance for each thesaurus based on the sorting results stored by the sorting-results-memory unit; a frequency-of-appearance-memory unit that stores the results calculated by the frequency-of appearance-calculation unit; a correlation-coefficient-calculation unit that calculates correlation coefficients between thesauruses; a correlation-coefficient-memory unit that stores the correlation coefficients between thesauruses that were calculated by the correlation-coefficient-calculation unit; a correlation-coefficient-total-calculation unit for each thesaurus that calculates the total of the correlation coefficients for each thesaurus; a correlation-coefficient-total-memory unit for each thesaurus that stores the total of the correlation coefficients for each thesaurus calculated by the correlation-coefficient-total-calculation unit for each thesaurus; and a graph-creation-display unit that creates and displays a graph based on the frequency of appearance stored by the frequency-of-appearance-memory unit and the correlation-coefficient totals for each thesaurus stored by the correlation-coefficient-total-memory unit for each thesaurus; and wherein the word-cutting unit and syntax-analysis unit perform the word-cutting process and syntax-analysis process again based on the thesauruses created by the thesaurus-creation unit.

The information-processing method of claim 2 comprises: an input step of inputting text data; a text-data-memory step of storing text data; a word-cutting step of executing a word-cutting process on text data; a syntax-analysis step of performing a syntax-analysis process on the text data on which the word-cutting process was performed; a thesaurus-creation step of creating thesauruses from the text data on which the syntax-analysis process was performed; a thesaurus-memory step of storing the thesauruses

created in the thesaurus-creation step; a word-cutting and syntax-analysis step of performing the word-cutting process and syntax-analysis process again based on the thesauruses stored in the thesaurus-memory step; a thesaurus-sorting step of performing a sorting process on the text data on which word-cutting and syntax-analysis were performed; a sorting-results-memory step of storing the sorting results from the thesaurus-sorting step; a frequency-of-appearance-calculation step of calculating the frequency of appearance for each thesaurus based on the sorting results stored in the sorting-results-memory step; a frequency-of-appearance-memory step of storing the results calculated in the frequency-of appearance-calculation step; a correlation-coefficient-calculation step of calculating correlation coefficients between thesauruses; a correlation-coefficient-memory step of storing the correlation coefficients between thesauruses that were calculated in the correlation-coefficient-calculation step; a correlation-coefficient-total-calculation step for each thesaurus of calculating the total of the correlation coefficients for each thesaurus; a correlation-coefficient-total-memory step for each thesaurus of storing the total of the correlation coefficients for each thesaurus calculated in the correlation-coefficient-total-calculation step for each thesaurus; and a graph-creation-display step of creating and displaying a graph based on the frequency of appearance stored in the frequency-of-appearance-memory step and the correlation-coefficient totals for each thesaurus stored in the correlation-coefficient-total-memory step for each thesaurus.

The information-processing program of claim 3 is executed on a computer to perform: an input step of inputting text data; a text-data-memory step of storing text data; a word-cutting step of executing a word-cutting process on text data; a syntax-analysis step of performing a syntax-analysis process on the text data on which the word-cutting process was performed; a thesaurus-creation step of creating thesauruses from the text data on which the syntax-analysis process was performed; a thesaurus-memory step of storing the thesauruses created in the thesaurus-creation step; a word-cutting and syntax-analysis step of performing the word-cutting

process and syntax-analysis process again based on the thesauruses stored in the thesaurus-memory step; a thesaurus-sorting step of performing a sorting process on the text data on which word-cutting and syntax-analysis were performed; a sorting-results-memory step of storing the sorting results from the thesaurus-sorting step; a frequency-of-appearance-calculation step of calculating the frequency of appearance for each thesaurus based on the sorting results stored in the sorting-results-memory step; a frequency-of-appearance-memory step of storing the results calculated in the frequency-of-appearance-calculation step; a correlation-coefficient-calculation step of calculating correlation coefficients between thesauruses; a correlation-coefficient-memory step of storing the correlation coefficients between thesauruses that were calculated in the correlation-coefficient-calculation step; a correlation-coefficient-total-calculation step for each thesaurus of calculating the total of the correlation coefficients for each thesaurus; a correlation-coefficient-total-memory step for each thesaurus of storing the total of the correlation coefficients for each thesaurus calculated in the correlation-coefficient-total-calculation step for each thesaurus; and a graph-creation-display step of creating and displaying a graph based on the frequency of appearance stored in the frequency-of-appearance-memory step and the correlation-coefficient totals for each thesaurus stored in the correlation-coefficient-total-memory step for each thesaurus.

[0007]
[Effect of the Invention]

The information-processing apparatus, information-processing method and information-processing program of this invention are such that: an input step inputs text data; a text-data-memory step stores text data; a word-cutting step executes a word-cutting process on text data; a syntax-analysis step performs a syntax-analysis process on the text data on which the word-cutting process was performed; a thesaurus-creation step creates thesauruses from the text data on which the syntax-analysis process was performed; a thesaurus-memory step stores the thesauruses created in the thesaurus-creation step; a word-cutting and syntax-analysis step performs

Our Ref : H1023555US01   Japan Patent Application No. 2002-360352

the word-cutting process and syntax-analysis process again based on the thesauruses stored in the thesaurus-memory step; a thesaurus-sorting step performs a sorting process on the text data on which word-cutting and syntax-analysis were performed; a sorting-results-memory step stores the sorting results from the thesaurus-sorting step; a frequency-of-appearance-calculation step calculates the frequency of appearance for each thesaurus based on the sorting results stored in the sorting-results-memory step; a frequency-of-appearance-memory step stores the results calculated in the frequency-of appearance-calculation step; a correlation-coefficient-calculation step calculates correlation coefficients between thesauruses; a correlation-coefficient-memory step stores the correlation coefficients between thesauruses that were calculated in the correlation-coefficient-calculation step; a correlation-coefficient-total-calculation step for each thesaurus calculates the total of the correlation coefficients for each thesaurus; a correlation-coefficient-total-memory step for each thesaurus stores the total of the correlation coefficients for each thesaurus calculated in the correlation-coefficient-total-calculation step for each thesaurus; and a graph-creation-display step creates and displays a graph based on the frequency of appearance stored in the frequency-of-appearance-memory step and the correlation-coefficient totals for each thesaurus stored in the correlation-coefficient-total-memory step for each thesaurus, so it is possible to extract the characteristics of text data based on the frequency of appearance of and correlation between thesauruses created from keywords extracted from the text data, and to analogize potential hidden meaning in the text data.

[0008]
[DESCRIPTION OF THE PREFERRED EMBODIMENT]
Fig. 1 is a block diagram showing the function of an embodiment of the information-processing apparatus of the invention. This embodiment comprises a personal computer or the like. As shown in the same figure, the embodiment is functionally constructed of the following blocks. The processing of each block is actually executed by a specified application

Our Ref : H1023555US01   Japan Patent Application No. 2002-360352

program, and each memory unit is made possible by a hard disc that is not shown in the figure.

[0009]

The function of each of the blocks will be briefly explained. The input unit 1 is used to input text data that is then stored in the text-memory 2. The word-cutting unit 3 executes a word-cutting process on the text data stored in the text-memory unit 2. The syntax-analysis unit 4 performs syntax analysis on the text data for which the word cutting process was performed.

[0010]

The thesaurus-creation unit 5 creates a thesaurus from the text data stored in the text-memory unit 2. The thesaurus-memory unit 6 stores the created thesaurus. The thesaurus-sorting unit 7 performs the sorting process on all samples for each thesaurus. The sorting-results-memory unit 8 stores the sorting results. The frequency-of-appearance-calculation unit 9 calculates the frequency of appearance in each thesaurus based on the data stored in the sorting-results-memory unit 8. The frequency-of-appearance-memory unit 10 stores the result calculated by the frequency-of-appearance-calculation unit 9.

[0011]

The correlation-coefficient-calculation unit 11 calculates the correlation coefficient between thesauruses. The correlation-coefficient-memory unit 12 stores the correlation coefficient calculated by the correlation-coefficient-calculation unit 11. The correlation-coefficient-total-calculation unit 13 for each thesaurus totals the found correlation coefficients for each thesaurus. The correlation-coefficient-total-memory unit 14 for each thesaurus stores the correlation-coefficient total for each thesaurus that was calculated by the correlation-coefficient-total-calculation unit 13 for each thesaurus. The graph-creation-display unit 15 creates and displays a graph based on the frequency of appearance stored in the frequency-of-appearance-memory unit

10 and the correlation-coefficient totals for each thesaurus that are stored in the correlation-coefficient-total-memory unit 14 for each thesaurus.

[0012]

Next, the processing procedure of this embodiment will be explained with reference to the flowchart shown in Fig. 2. Here, the case of analyzing text data of complaints or demands from customers will be explained. First, in step S1, text data from each customer is input from the input unit 1. For example, suppose that the complaint from one customer, 'Last week, I ordered a part, but it has not been delivered yet.' is input. The input text data is stored in the text-memory unit 2.

[0013]

Next, in step S2, the word-cutting unit 3 performs a word-cutting process using a specified text-mining tool (application software). For example, the text above becomes 'Last week I ordered a part, but it has not been delivered yet.'

[0014]

Next, in step S3, the syntax-analysis unit 4 performs a syntax-analysis process using a text-mining tool. For example, the text above becomes 'Last week I ordered a part but, it has not been delivered yet.'

[0015]

Next, in step S4, the thesaurus-creation unit 5, creates a thesaurus with a collection of synonyms (keywords). For example, as shown in Fig. 3, synonyms or keywords such as 'one week ' are collected in the 'last week' thesaurus. Also, keywords such as 'ordered, but' are collected in the 'order' thesaurus. Also, keywords such as 'deliver' are collected in the 'deliver' thesaurus. Also, keywords such as 'parts' are collected in the 'parts' thesaurus. Also, keywords such as 'information' are collected in the 'contact' thesaurus. The created thesauruses are stored in the thesaurus-memory unit 6.

[0016]

Next, in step S5, the word-cutting unit 3 performs the word-cutting process again based on the thesauruses created now and stored in the thesaurus-memory unit 6, and the syntax-analysis unit 4 performs the syntax-analysis process again.

[0017]

Next, in step S6, the thesaurus-sorting unit 7 performs sorting of the contents of the text data from all customers for each thesaurus. For example, '1' is set for all of the thesauruses that are contained in text data such as a complaint from a user, for each user, and sets '0' for the thesauruses that are not contained. The sorting results are stored in the sorting-results-memory unit 8.

[0018]

Fig. 4 shows the sorting results that are stored in the sorting-results-memory unit 8. In the figure, 'K-1', 'K-2', 'K-3', ... 'K-n' indicate ID numbers that identify the customer. In this example, it can be seen that text data containing keywords contained in the 'order' and 'parts' thesauruses was input by customer K-1.

[0019]

Next, in step S7, the correlation-coefficient-calculation unit 11 finds correlation coefficients between thesauruses. For example, the correlation coefficient for 'order' and 'deliver' is expressed by the following equation.

[0020]

Correlation coefficient $r_{order \cdot delivery} = (S_{order \cdot delivery})/(S_{order} \cdot S_{delivery})$

Where $S_{order \cdot delivery}$ is the covariance, and S_{order} and $S_{delivery}$ are each standard deviations.

[0021]

Our Ref : H1023555US01   Japan Patent Application No. 2002-360352

Covariance $S_{order-delivery}$ = ((order$_1$ - order)(delivery$_1$ - delivery) + (order$_2$ - order)(delivery$_2$ - delivery) + ... + (order$_n$ - order)(delivery$_n$ - delivery))/(n - 1)

[0022]

Standard deviation S_{order} = ((order$_1$ - order)^2 + (order$_2$ - order)^2 + ... + (order n - order)^2)/(n - 1)

Standard deviation $S_{delivery}$ = ((delivery$_1$ - delivery)^2 + (delivery$_2$ - delivery)^2 + ... + (delivery n - delivery)^2)/(n - 1)

[0023]

Similarly, correlation coefficients are found between all of the thesauruses, and stored in the correlation-coefficient-memory unit 12. Fig. 5 shows the correlation coefficients between the thesauruses. For example, the correlation coefficient between the thesaurus 'last week' and the thesaurus 'order' is 0.025. The correlation coefficient between identical thesauruses is 1.

[0024]

Next, in step S8, the correlation-coefficient-total-calculation unit 13 for each thesaurus totals the correlation coefficients stored in the correlation-coefficient-memory unit 12 for each thesaurus. For example, in the case of the thesaurus 'last week', the total is 1 + 0.025 + 0.038 + 0.001 + ... When doing this, the correlation coefficient 1 between identical thesauruses is omitted. Similarly, the correlation-coefficient totals are found for the other thesauruses such as 'order', 'deliver', 'parts', etc. The correlation-coefficient totals that were found for each of the thesauruses are stored in the correlation-coefficient-total-memory unit 14 for each thesaurus.

[0025]

Next, in step S9, the frequency-of-appearance-calculation unit 9 finds the frequency of appearance for each thesaurus. That is, as shown in Fig. 6, the frequency of appearance for each thesaurus is found based on the sorting

results (Fig. 4) for each thesaurus. In the example of Fig. 6, for example, for the thesaurus 'last week', it is seen that that thesaurus 'last week' is contained in the text data of the complaints from customers K-2, K-3, ... K-n. By totaling the values for each customer for the thesaurus 'last week', the number of appearances A is calculated. Similarly, the number of appearances for thesaurus 'order' is B, the number of appearances for thesaurus 'deliver' is C and the number of appearances for thesaurus 'parts' is D. The total number of appearances for all of the thesauruses $\Sigma(A + B + C + D + ...)$ is found, and the frequency of appearance for each thesaurus is expressed as a percentage.

[0026]

For example, the frequency of appearance of the thesaurus 'last week' is $(A/\Sigma(A + B + C + D + ...)) * 100(\%)$. The calculated frequencies of appearance for each of thesauruses are stored in the frequency-of-appearance-memory unit 10.

[0027]

Next, in step S10, the graph-creation-display unit 15 plots the frequency of appearance (%) for each thesaurus along the x-axis, and the correlation coefficient totals for each thesaurus along the y-axis, to create a graph. Fig. 7 shows the created graph.

[0028]

As shown in Fig. 7, in the complaints from customers, thesauruses whose number of appearances is not so large but whose connection (some kind of relationship) with other thesauruses is large appear in the first group.

[0029]

Also, thesauruses that do not have such a strong connection with other thesauruses but whose number of appearances is large, or in other words, thesauruses that cannot be ignored because they are mentioned frequently,

appear in the third group.

[0030]

Here, significance is not found based on the size of the value of the correlation coefficient, but a fixed level is set as a reference, and it is determined that there is a strong connection when the correlation exceeds that reference level, and it is determined that the connection is weak when the correlation is below that level.

[0031]

Whether or not the y-coordinate value exceed a fixed level is important, and for thesauruses that exceed a fixed level, there is a high possibility that there is some significance when connected with other keywords. In this case, since the text is a complaint from a customer concerning a part, this 'significance' is a complaint, or in other words, it can be analogized that it indicates a 'potential dissatisfaction'.

[0032]

The construction and operation of the embodiment described above are examples, and needless to say, can be suitably changed within a range that does not deviate from the object of the invention.

[0033]

[BRIEF DESCRIPTION OF THE DRAWINGS]

Fig. 1 is a block diagram showing the function of a first embodiment of the information-processing apparatus of the invention.

Fig. 2 is a flowchart for explaining the processing procedure of an embodiment of the invention.

Fig. 3 is a drawing showing an example of a thesaurus having a collection of synonyms.

Fig. 4 is a drawing showing the sorting results for each thesaurus.

Fig. 5 is a drawing showing the correlation coefficients for each thesaurus.

Our Ref : H1023555US01   Japan Patent Application No. 2002-360352

Fig. 6 is a drawing showing the frequency of appearance of each thesaurus.

Fig. 7 is a graph showing the relationship between the frequency of appearance and the correlation coefficients of each thesaurus.

[Description of Reference Numbers]

1 Input unit

2 Text-memory unit

3 Word-cutting unit

4 Syntax-analysis unit

5 Thesaurus-creation unit

6 Thesaurus-memory unit

7 Thesaurus-sorting unit

8 Sorting-results-memory unit

9 Frequency-of-appearance-calculation unit

10 Frequency-of-appearance-memory unit

11 Correlation-coefficient-calculation unit

12 Correlation-coefficient-memory unit

13 Correlation-coefficient-total-calculation unit for each thesaurus

14 Correlation-coefficient-total-memory unit for each thesaurus

15 Graph-creation-display unit

[ABSTRACT]

[Object]   The invention makes it possible to detect the characteristics of text data, and to analogize potential hidden meaning in the text data.

[Means to Solve]   A word-cutting unit 3 performs a word-cutting process on the text data input from the input unit 1, a syntax-analysis unit 4 performs syntax analysis and a thesaurus-creation unit 5 creates thesauruses from the results, then after performing word cutting and syntax analysis again, a thesaurus-sorting unit 7 performs sorting, and a frequency-of-appearance unit calculates the frequency of appearance of the thesauruses, a correlation-coefficient-calculation unit 11 calculates